Exploratory Data Analysis (EDA) on the Titanic Dataset

1. Introduction

This report presents an exploratory data analysis (EDA) of the Titanic dataset using Python libraries: Pandas, Matplotlib, and Seaborn.

The main objective is to discover trends, patterns, and anomalies by visual and statistical exploration. We investigate factors influencing passenger survival, analyze distributions, relationships between variables, and summarize the insights.

2. Dataset Overview

The Titanic dataset contains information about passengers aboard the RMS Titanic. It includes demographic, class, fare, and survival status.

Key Features:

- Survived (0 = No, 1 = Yes)
- Pclass (Ticket class: 1 = 1st, 2 = 2nd, 3 = 3rd)
- Sex
- Age
- SibSp (Siblings/Spouses aboard)
- Parch (Parents/Children aboard)
- Fare
- Embarked (Port of Embarkation)

Initial inspection revealed missing values in 'age', 'embarked', and 'deck'.

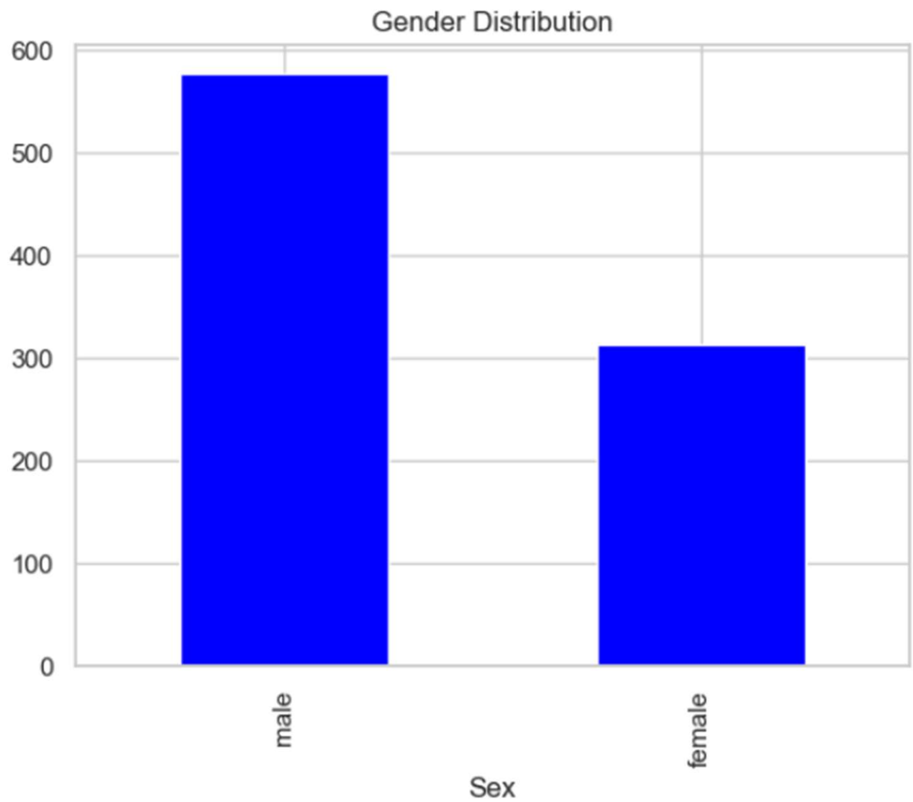
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Toggle output scrolling

3. Univariate Analysis

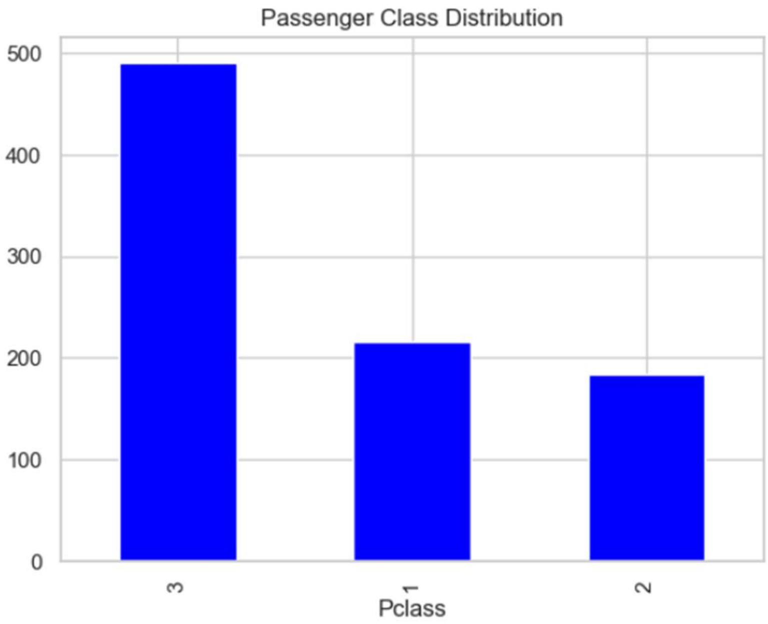
3.1 Gender Distribution



Observation:

The dataset contains more male passengers than female. Males: ~65%, Females: ~35%.

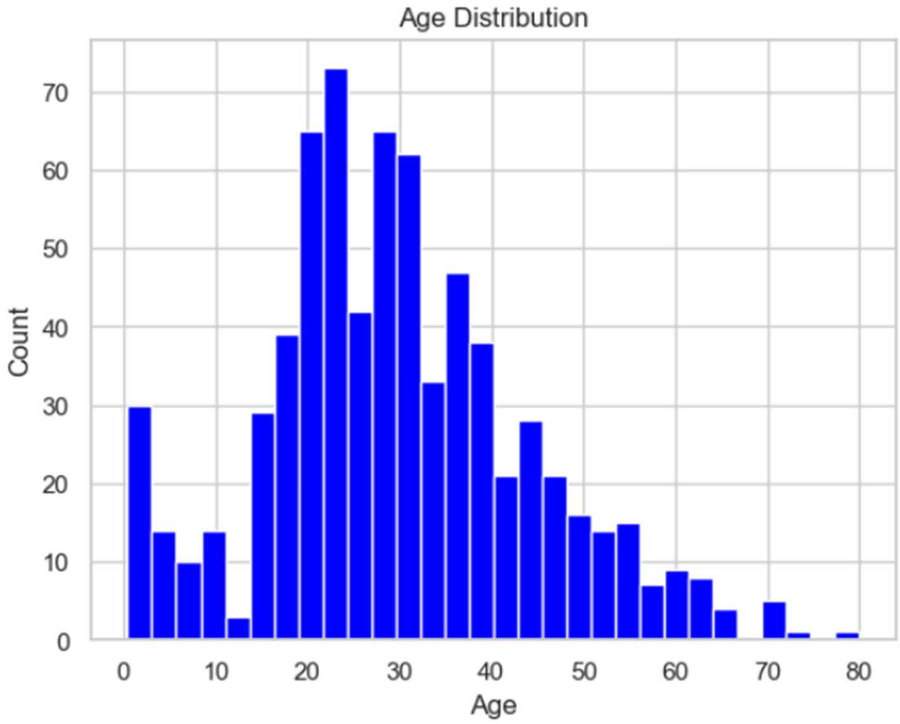
3.2 Passenger Class Distribution



Observation:

Third class had the highest number of passengers, followed by First and Second classes.

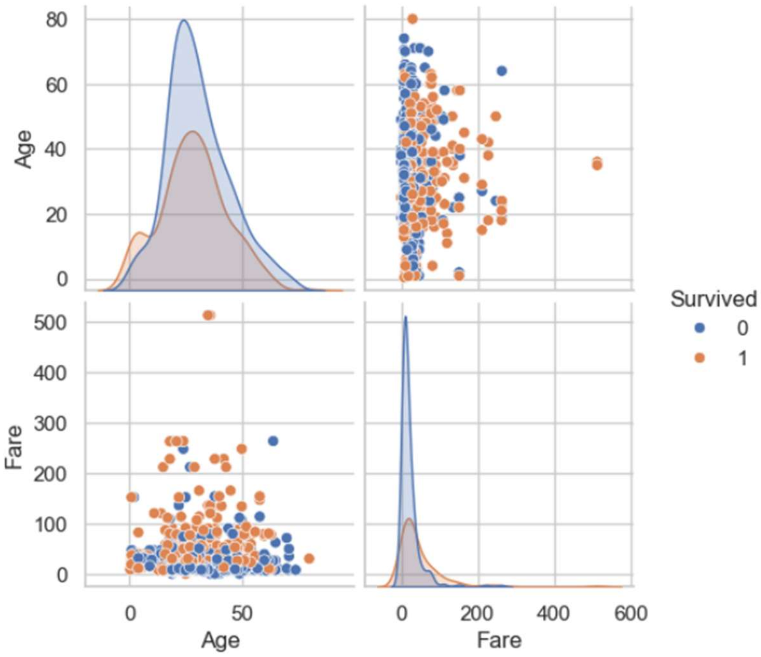
3.3 Age Distribution



Observation:

Majority of passengers were aged 20–40. Some missing values exist in the 'age' column

3.4 Fare Boxplot

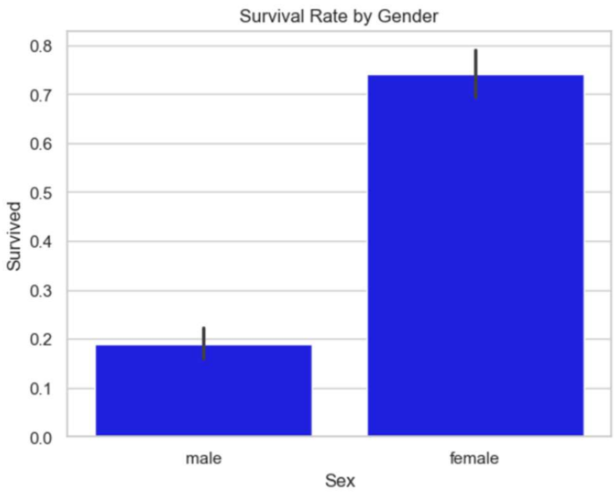


Observation:

Fare distribution is right-skewed. Some passengers paid extremely high fares, indicating luxury services.

4. Bivariate & Multivariate Analysis

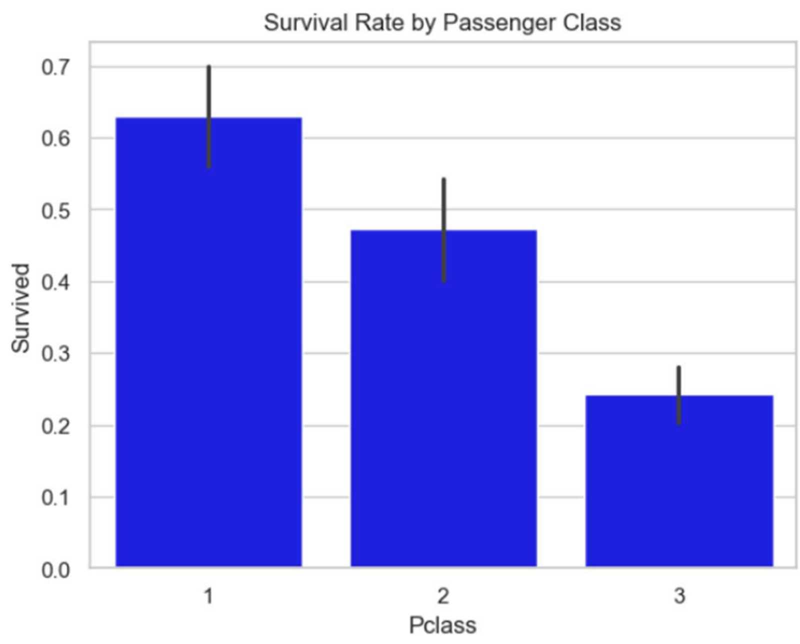
4.1 Survival Rate by Gender



Observation:

Females had a significantly higher survival rate (~75%) than males (~20%).

4.2 Survival Rate by Passenger Class



Observation:

First-class passengers had a much higher survival rate than third-class passengers.

5. Key Insights & Summary

- Female passengers and those in first class were more likely to survive.
- Younger passengers had slightly better chances of survival.
- Fare prices vary widely; higher fares correlate slightly with survival.
- Missing values in 'age' and 'embarked' should be handled before modeling.
- Class, sex, and fare are the most influential features in predicting survival.